A systematic review highlights the need to investigate the content validity of patient-reported outcome measures for physical functioning in patients with low back pain

Chiarotto, Alessandro; Ostelo, Raymond W.; Boers, Maarten; Terwee, Caroline B.

published in
Journal of Clinical Epidemiology
2018

DOI (link to publisher)
10.1016/j.jclinepi.2017.11.005

document version
Publisher's PDF, also known as Version of record

Journal of Clinical Epidemiology 95 (2018) 73–93

Journal of Clinical Epidemiology

REVIEW

A systematic review highlights the need to investigate the content validity of patient-reported outcome measures for physical functioning in patients with low back pain

Alessandro Chiarotto[a,b,*], Raymond W. Ostelo[a,b], Maarten Boers[b,c], Caroline B. Terwee[b]

[a]*Department of Health Sciences, Amsterdam Movement Sciences Research Institute, Vrije Universiteit, Amsterdam, The Netherlands*
[b]*Department of Epidemiology and Biostatistics, Amsterdam Public Health Research Institute, VU University Medical Center, Amsterdam, The Netherlands*
[c]*Amsterdam Rheumatology and Immunology Center, VU University Medical Center, Amsterdam, The Netherlands*

Accepted 8 November 2017; Published online 14 November 2017

Abstract

Objectives: To summarize the evidence on content and structural validity of 17 patient-reported outcome measures (PROMs) to measure physical functioning in patients with low back pain (LBP).

Study Design and Setting: MEDLINE, EMBASE, CINAHL, PsycINFO, SportDiscus, and Google Scholar were searched (February 2017). Records on development and studies assessing content validity or unidimensionality in patients with LBP were included. Two reviewers defined eligible studies and assessed their methodological quality with updated Consensus-based Standards for the Selection of Health Measurement Instruments standards. Evidence was synthesized for three separate aspects of content validity: relevance, comprehensiveness, and comprehensibility, and for unidimensionality, a modified GRADE approach was applied to evidence synthesis.

Results: High-quality evidence showed that 24-item Roland Morris Disability Questionnaire (RMDQ-24) is a comprehensible but not comprehensive PROM. Low to very low quality evidence underpinned the content validity of the other PROMs. Unidimensionality was: sufficient for Brief Pain Inventory pain interference subscale (moderate quality evidence); inconsistent for RMDQ-23, Oswestry Disability Index 2.1a (ODI 2.1a), and Quebec Back Pain Disability Scale (moderate quality); insufficient for RMDQ-24, ODI 1.0, and RMDQ-18 (high quality) and Short Form 36 physical functioning subscale (SF36-PF, moderate quality).

Conclusion: The content validity of PROMs to measure physical functioning in patients with LBP is understudied. Structural validity of several widely used PROMs is problematic.

Keywords: Patient-reported outcome measures; Physical functioning; Low back pain; Content validity; Unidimensionality; COSMIN

1. Introduction

Low back pain (LBP) is a burdensome and costly health condition that affects many individuals and health care systems [1,2]. Thus, measurement of its impact on patients is important in clinical research and practice [3]. Physical functioning is considered by researchers, clinicians and patients to be the most important outcome domain to measure in LBP clinical trials [4]. Most frequently, patient-reported outcome measures (PROMs) are used to measure this domain, especially the Oswestry Disability Index (ODI) and the Roland Morris Disability Questionnaire (RMDQ) [5,6]; these two measurement instruments have also been recommended by international standardization initiatives [7–11].

The choice of an adequate instrument is strongly determined by its validity, that is, the extent to which it accurately measures what is supposed to measure [12]. The Consensus-based Standards for the Selection of Health Measurement Instruments (COSMIN) taxonomy distinguished five subdomains of validity [13], among which content validity is the first one to be considered when

Conflict of interest: The authors of this manuscript declare that they do not have any conflict of interest related to the content of this manuscript.

Financial support: The authors of this manuscript would like to acknowledge the EUROSPINE Task Force Research for providing funding for this study (EUROSPINE TFR 5-2015). This funding body did not have any role in designing the study, in collecting, analyzing and interpreting the data, in writing this manuscript, and in deciding to submit it for publication.

* Corresponding author. Department of Health Sciences, Faculty of Science, Amsterdam Movement Sciences Research Institute, Amsterdam Public Health Research Institute, Vrije Universiteit, de Boelelaan 1085, Room U-601, 1081 HV, Amsterdam, The Netherlands.

E-mail address: a.chiarotto@vu.nl (A. Chiarotto).

https://doi.org/10.1016/j.jclinepi.2017.11.005
0895-4356/

What is new?

Key findings

- The quality of evidence on content validity of most patient-reported outcome measures (PROMs) to measure physical functioning in patients with low back pain (LBP) is insufficient to draw any firm conclusion about this measurement property.
- High quality evidence suggests that RMDQ-24, RMDQ-18, and ODI 1.0 are not unidimensional tools. Less robust evidence suggests BPI-PI is unidimensional and SF36-PF is not; for RMDQ-23, ODI 2.1a, MPI-PI, and QBPDS results are inconsistent.

What this adds to what is known?

- This is the first systematic review to thoroughly assess the content validity of these widely used PROMs.
- Our findings do not support the use of total scores of RMDQ-24, ODI 1.0, RMDQ-18, and SF36-PF and cast serious doubt on the use of the total scores of RMDQ-23, ODI 2.1a, MPI-PI, and QBPDS.

What is the implication and what should change now?

- All included PROMs urgently require thorough assessment of content validity through qualitative research with patients to explore their relevance, comprehensiveness, and comprehensibility for measuring physical functioning in patients with LBP. Head-to-head comparisons of different PROMs would be useful.
- Unidimensionality of various PROMs needs to be better investigated, and the impact of multidimensionality can be documented with bifactor analysis or multidimensional item response theory to determine the most appropriate dimensional structure.

selecting a PROM [14]. Content validity refers to ''the degree to which the content of an instrument is an adequate reflection of the construct to be measured'' [13]; it deals with the relevance, comprehensiveness, and comprehensibility of a PROM with respect to construct, target population, and context of use of interest [15−17]. Content validity influences all other measurement properties. For example, irrelevant items can lead to poor internal consistency, unidimensionality, and interpretability of a PROM, and a lack of comprehensiveness (i.e., absence of key aspects in an instrument) can reduce responsiveness (Terwee et al., 2017, unpublished data).

Next in importance is structural validity, which refers to ''the degree to which the scores of an instrument are an adequate reflection of the dimensionality of the construct to be measured'' [14]. Physical functioning is usually considered to be a broad but unidimensional domain. For example, in the Patient-Reported Outcomes Measurement Information System (PROMIS) conceptual framework, it was defined as ''one's ability to carry out various activities that require physical capability, ranging from self-care to more vigorous activities'' [18]. In research and practice, the total score of ODI and RMDQ is routinely used, under the implicit assumption that these instruments measure one single domain [19]. Therefore, a PROM selected to measure physical functioning in patients with LBP is expected, first, to have good content validity and, second, to be unidimensional.

Two recent systematic reviews found limited evidence for good content validity and moderate evidence for unidimensionality of the Quebec Back Pain Disability Scale (QBPDS), another well-known and used PROM in LBP [20], and a lack of head-to-head comparisons of content and structural validity between the ODI (version) 2.1a vs. the 24-item RMDQ (RMDQ-24) in patients with LBP [21]. No systematic reviews are available on content and structural validity of ODI, RMDQ, or of other PROMs recommended to measure physical functioning in patients with LBP, such as the PROMIS Physical Function 4-item short form (PROMIS-PF-4) recommended by the National Institutes of Health (NIH) Task Force for research standards in chronic LBP [22].

Any systematic review of PROM content validity should not only include content validity studies but also the original PROM development study and the content of the instrument itself. The COSMIN initiative has recently developed methodological guidance for this type of reviews, with criteria to determine what constitutes sufficient content validity, and a method to integrate methodological quality and results into an evidence synthesis rating system (Terwee et al., 2017, unpublished data). The COSMIN checklist and review methods for other measurement properties (including structural validity) have also been updated (Mokkink et al., 2017 and Prinsen et al., 2017, unpublished data).

The present study applies this COSMIN methodology to systematically review content and structural validity of a set of PROMs to measure physical functioning in patients with LBP [5−7,9−11,22]. This review is embedded within an international multidisciplinary collaboration to develop a core outcome measurement set [23,24] for clinical trials in patients with nonspecific LBP (nsLBP) [4].

2. Methods

This systematic review was conducted and reported according to the Preferred Reporting Items for Systematic

Reviews and meta-Analysis statement [25]. A protocol was written a priori and registered in the international prospective register of systematic reviews, accessible at: http://www.crd.yor.ac.uk/PROSPERO/ (registration number: 42015019840).

In a previous study to find consensus on core outcome domains for clinical trials in patients with nsLBP, physical functioning was selected as a core domain with the following definition: "the patient's ability to carry out daily physical activities required to meet basic needs, ranging from self-care to more complex activities that require a combination of skills" [4]. This definition was the starting point for this systematic review, which aimed to (1) assess and compare the content validity of various PROMs to measure this domain in patients with nsLBP and (2) assess and compare the structural validity of the same PROMs in patients with nsLBP.

Seventeen PROMs were selected as potential core outcome measurement instruments for physical functioning in patients with LBP, including those recommended [7–11,22] and those most widely used [5,6] (Table 1). This selection led to the inclusion of both disease-specific (e.g., ODI 2.1a, RMDQ-24, and QBPDS) and generic-specific instruments (e.g., BPI-PI, MPI-PI, SF36-PF, and PROMIS-PF short forms). The assumption that all these PROMs measure the same domain in patients with nsLBP underlain this review, as, for instance, they are often statistically pooled in meta-analysis of clinical trials [26,27]. For structural validity, the unidimensionality of the total score of each PROM was assessed, as the use of such scores has been advocated by their developers (Table 1), and subsequently, these scores have been routinely used in clinical research and practice.

2.1. Data sources and searches

MEDLINE (through the interface PubMed), EMBASE (Embase.com), CINAHL (EBSCOhost), PsycINFO (EBSCOhost), and SportDiscus (EBSCOhost) were last searched on February 6, 2017. The search strategy consisted of three groups of search terms combined with the Boolean operator 'AND', representing the following components: (1) the names of the PROMs, (2) LBP, and (3) measurement properties. A previously developed search filter retrieved studies on the measurement properties in PubMed [28]; the same filter was adapted for all the other databases (Appendix A). No restrictions for language or time were adopted in the search strategies. Google Scholar was also searched (last on February 13, 2017) with the full names of the PROMs, and the first 100 hits for each PROM were screened for inclusion. Citation tracking of the eligible studies was carried out by consulting the database Web of Science and by checking their references.

2.2. Study selection

The studies presenting the development of the 17 PROMs (Table 1) were included for the assessment of content validity, irrespective of the format in which they were presented (e.g., journal article, book chapter, and user guide). Two of these PROMs (i.e., Pain Interference subscale of the Brief Pain Inventory and pain interference items of the Multidimensional Pain Inventory) were not developed to measure the construct physical functioning but were included because they had been recommended by the Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT) for this domain [11]. Content validity studies were eligible if they were full-text original articles, about adult patients with nsLBP or professionals (e.g., researchers, clinicians) to assess the relevance, comprehensiveness, or comprehensibility of the content of at least one of the PROMs. Studies on cross-cultural adaptation of the PROMs were included as content validity studies if they performed a pretest of the adapted questionnaire (see guidance for cross-cultural adaptations: [29]), in which its comprehensibility was assessed in patients with nsLBP.

Structural validity studies were eligible if they were full-text original articles about adult patients with nsLBP and assessing the dimensionality of one or more of the PROMs with (bi)factor analysis or item response theory (IRT) analysis. A study including patients with specific LBP or a mixed population of patients with pain was included if at least 75% of the patients were classified as having nsLBP. Studies had to apply the original version of each of the selected PROMs, except for ODI 2.1a in which all three versions of ODI 2 (i.e., 2.0, 2.1, and 2.1a) were considered the same, given the very minor grammatical adjustments made between versions [30].

Inclusion criteria were applied by two reviewers (A.C. and R.W.O.) independently to titles and abstracts of the hits retrieved in the databases. Subsequently, potentially eligible full texts were screened independently by the same two reviewers. Consensus on inclusion was sought between reviewers, and in case of disagreement, a third reviewer (C.B.T.) made decisions.

2.3. Quality assessment and data extraction

The methodological quality of a PROM development was assessed using COSMIN standards. These comprise 35 items subdivided in two parts: one addressing the concept elicitation study performed with patients to identify relevant items for a new PROM (including a clear description of the construct and a theory or conceptual framework from which it originates) and the other addressing the cognitive interview study performed with patients to evaluate comprehensiveness and comprehensibility (Terwee et al., 2017, unpublished data). A second set of newly developed COSMIN standards was used to assess the methodological quality of the studies on content validity. These include 38 items subdivided in two parts: one addressing studies that query patients about relevance, comprehensiveness, and comprehensibility, and the other

Table 1. Patient-reported outcome measures (PROM) selected as potential core outcome measurement instruments to measure physical functioning in clinical trials in patients with low back pain

PROM	Name abbreviation	Characteristics				Recommended by standardization initiatives for LBP or chronic pain
		Recall period	Number of items	Response options	Total score range	
Oswestry disability index, version 1.0	ODI 1.0	Undefined	10	0–5 rating scale	0–100	Original LBP core set [7,10]
Oswestry disability index, version 2.1a	ODI 2.1a	Undefined	10	0–5 rating scale	0–100	Original LBP core set [7,10]; ICHOM standard set for LBP [9]
Low back pain disability questionnaire—chiropractic version	CLBPDQ	Undefined	10	0–5 rating scale	0–100	
Low back pain disability questionnaire—modified version	MLBPDQ	Today	10	0–5 rating scale	0–100	
Roland Morris disability questionnaire—24-item	RMDQ-24	Today	24	0–1 yes/no	0–24	Original LBP core set [7,10]
Roland Morris disability questionnaire—23-item	RMDQ-23	Today	23	0–1 yes/no	0–23	Original LBP core set [7,10]
Roland Morris disability questionnaire—18-item	RMDQ-18	Today	18	0–1 yes/no	0–18	Original LBP core set [7,10]
Brief pain inventory—pain interference subscale	BPI-PI	Last 24 h	7	0–10 numeric scale	0–10	IMMPACT for chronic pain trials [11]
Multidimensional pain inventory—pain interference items	MPI-PI	Undefined	9	0–6 rating scale	0–6	IMMPACT for chronic pain trials [11]
Short Form Health Survey 36—physical functioning subscale	SF36-PF	Now	10	1–3 rating scale	0–100	
Low back pain rating scale—disability index	LBPRS-DI	Undefined	15	0–2 rating scale	0–30	
Quebec back pain disability scale	QBPDS	Today	20	0–5 rating scale	0–80	
Patient-reported outcomes measurement information system physical function short form—4-item	PROMIS-PF-4	Undefined	4	1–5 rating scale	0–100[a]	NIH Task Force for chronic LBP [22]
Patient-reported outcomes measurement information system physical function short form—6-item	PROMIS-PF-6	Undefined	6	1–5 rating scale	0–100[a]	
Patient-reported outcomes measurement information system physical function short form—8-item	PROMIS-PF-8	Undefined	8	1–5 rating scale	0–100[a]	
Patient-reported outcomes measurement information system physical function short form—10-item	PROMIS-PF-10	Undefined	10	1–5 rating scale	0–100[a]	
Patient-reported outcomes measurement information system physical function short form—20-item	PROMIS-PF-20	Undefined	20	1–5 rating scale	0–100[a]	

Abbreviations: PROM, patient-reported outcome measures; LBP, low back pain; NIH, National Institutes of Health.
[a] This range is expressed in *t* scores with a population mean of 50 and a standard deviation of 10.

addressing studies that query professionals about relevance and comprehensiveness (Terwee et al., 2017, unpublished data). Each standard is scored on a 4-point rating scale, that is, ''very good'', ''adequate'', ''doubtful'', and ''inadequate''. Total scores are determined for the two parts of the development study (concept elicitation and cognitive interview) separately, as well as for each aspect of a content validity study separately (i.e., relevance, comprehensiveness, and comprehensibility). Studies with patients or professionals are also rated separately. A total score per box or part of a box is obtained by taking the lowest rating of any item in the (part of the) box (i.e., worst score

counts) [31]. More detailed information on these new standards can be found elsewhere (Terwee et al., 2017, unpublished data).

We assessed the methodological quality of the studies on structural validity with the newly developed COSMIN risk of bias checklist for PROMs (Mokkink et al., 2017, unpublished data). The 4-point rating scale and worst score counts method are the same as those for content validity. For both measurement properties, two reviewers (A.C. and C.B.T.) assessed the quality separately and determined the consensus ratings in a face-to-face meeting. Information extracted included the construct to be measured, target population, and context of use (PROM development studies); patient characteristics (concept elicitation and cognitive interview studies; validity studies); and results (validity studies). Data were extracted by one reviewer (A.C.); a second reviewer (R.W.O.) double-checked the accuracy of a random 25% of the extracted information, stratified for each PROM.

2.4. Evidence synthesis

Evidence synthesis comprised two steps. First, the results of PROM development and content validity studies were rated by two reviewers (A.C. and C.B.T.) independently according to 10 established criteria: five on relevance, one on comprehensiveness, and four on comprehensibility (Terwee et al., 2017, unpublished data). Each criterion could be scored as positive (+), negative (−), or indeterminate (?). The same criteria were also scored based on the content of the PROM itself (Terwee et al., 2017, unpublished data). An overall sufficient (+), insufficient (−), or inconsistent (±) score was provided for relevance, comprehensiveness, and comprehensibility of each PROM, by jointly assessing all the results and reviewer ratings on the same PROM (irrespective of language and country). The study results were rated according to a modified version of the consensus-based criteria proposed by Prinsen et al. [14]; these criteria were amended

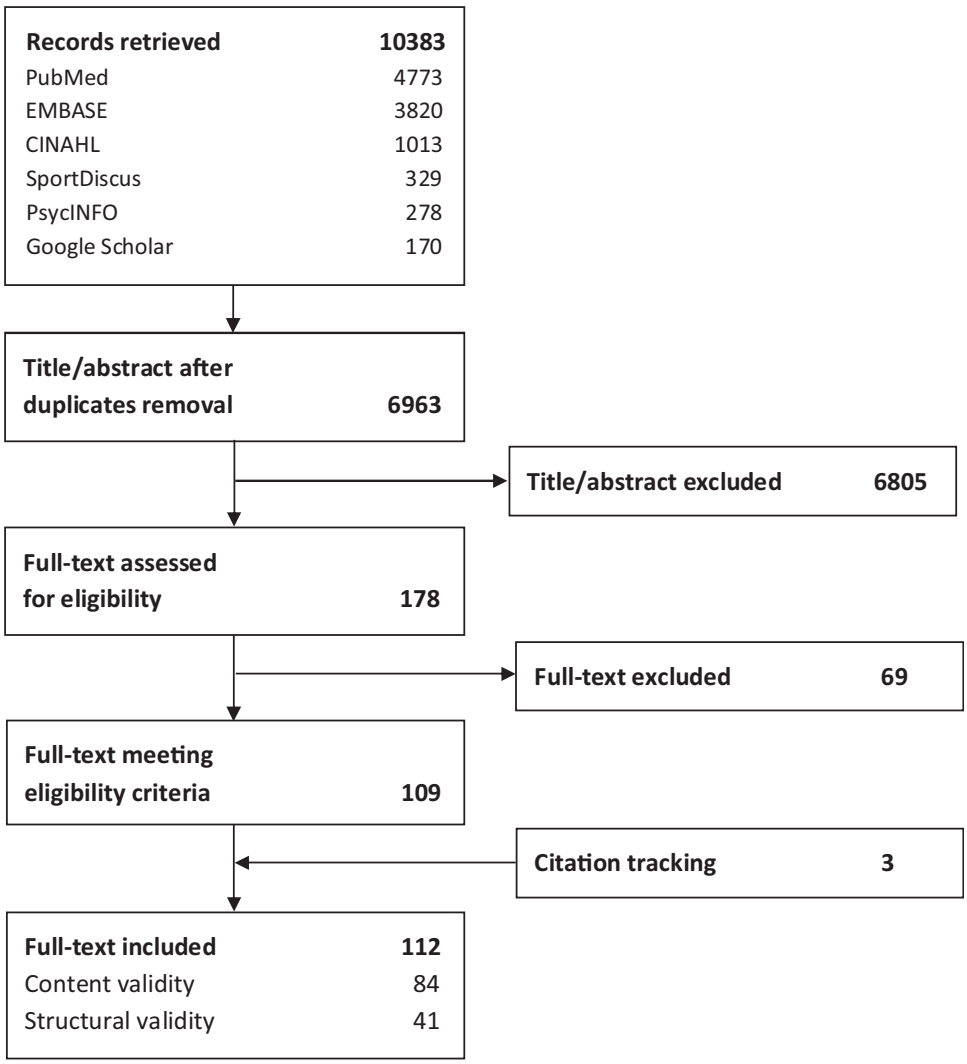

Fig. 1. Flow chart of results of search strategy and selection of records.

Table 2. Characteristics and quality assessment of the studies on the development of the included patient-reported outcome measures (PROM)

PROM	Reference	Primary language	Construct definition
ODI 1.0	Fairbank 1980 [39]	English (UK)	Disability = the limitations of a patient's performance compared with that of a fit person
ODI 2.1a	Fairbank 1980, Meade 1986, Baker 1989 [32,39,45]	English (UK)	Probably like ODI 1.0
CLBPDQ	Fairbank 1980, Hudson-Cook 1989 [39,41]	English (UK)	Probably like ODI 1.0
MLBPDQ	Fairbank 1980, Fritz 2001 [39,40]	English (US)	Probably like ODI 1.0
RMDQ-24	Roland 1983 [47]	English (UK)	Self-rated disability due to LBP referring to a range of aspects of daily living
RMDQ-23	Roland 1983, Patrick 1995 [46,47]	English (UK)	Back pain-specific functional status
RMDQ-18	Roland 1983, Stratford 1997 [47,50]	English (Canada)	Probably like RMDQ-24
BPI-PI	Daut 1983, Cleeland 1994, Cleeland 2009 [35–37]	English (US)	Sensory dimension of pain (intensity, or severity) and the 'reactive' dimension of pain (interference with daily function)
MPI-PI	Kerns 1985 [42]	English (US)	First section: evaluation of perceived pain intensity and the impact of pain on various aspects of the patients' lives
PF-SF36	Stewart 1992, Ware 1992 [49,51]	English (US)	Performance of or capacity to perform a variety of physical activities normal for people in good health. (…) those physical activities most likely to be the same for all people regardless of their life situation
LBPRS-DI	Manniche 1994 [44]	Danish	Daily tasks (…). The aim is to record both physical and psychological functional loss
QBPDS	Kopec 1996 [43]	English, French (Canada)	Disability = difficulty experienced while performing simple tasks. Complex tasks were avoided because many simple activities are necessary to perform those tasks
PROMIS-PF-4, PROMIS-PF-6, PROMIS-PF-8, PROMIS-PF-10, PROMIS-PF-20	Cella 2007, DeWalt 2007, Bruce 2009, Cella 2010, Rose 2014, PROMIS scientific standards [18,31,33,34,38,48]	English (US)	Physical function latent trait = ability to carry out various activities that require physical capability, ranging from self-care (basic activities of daily living [ADL]) to more vigorous activities that require increasing degrees of mobility, strength, or endurance

Abbreviations: LBP, low back pain; ?, not reported; /, not applicable.

to allow the assessment of unidimensionality, irrespective of the statistical techniques used (Appendix B). More specific information on how to apply these criteria is provided elsewhere (Terwee et al., 2017 and Prinsen et al., 2017, unpublished data).

Second, the quality of evidence was rated according to GRADE [32], adapted for this type of review, into "high", "moderate", "low", or "very low", taking into account the study quality, consistency of results across studies, and reviewers' ratings (for content validity only) (Terwee et al., 2017 and Prinsen et al., 2017, unpublished data).

3. Results

From more than 10,000 initial records, 171 were retrieved for full-text assessment, and 112 were selected (Figure 1). Sixty-nine records proved ineligible: 24 included patients with nsLBP within an heterogeneous clinical population without providing results for the nsLBP group separately, 14 assessed other measurement properties, eight performed a cross-cultural adaptation of a PROM without a pretest of the adapted version, seven were not in nsLBP, three did not assess measurement properties, three

Target population	Intended context of use	Concept elicitation study	
		COSMIN quality rating	Were patients involved?
LBP	Assessment of patients with LBP: guide to a patient's treatment program	Inadequate	No
Probably like ODI 1.0	Probably like ODI 1.0	Inadequate	No
Probably like ODI 1.0	Probably like ODI 1.0	Inadequate	No
Probably like ODI 1.0	Probably like ODI 1.0	Inadequate	No
LBP	Outcome measure for LBP clinical trials	Inadequate	No
LBP, including sciatica or leg pain	Probably like RMDQ-24	Inadequate	No
Probably like RMDQ-24	Probably like RMDQ-24	Inadequate	No
Cancer pain	Self-report measures of cancer pain, application to studies of pain and its treatment in the United States and internationally	Doubtful	Yes
Chronic pain	Brief but comprehensive assessment of the subjective experience of pain for inclusion as part of an extended assessment protocol	Inadequate	No
General and patient populations	Clinical practice and research, healthy policy evaluations, and general population surveys	Inadequate	No
Acute or chronic LBP	Compact, readily usable and simultaneously complete indirect measurement of low back pain, primarily for use in clinical trials, but also as a status assessment in clinical practice.	Inadequate	No
Back pain	Multipurpose questionnaire for clinical trials, and for patients participating in treatment or rehabilitation programs	Doubtful	Yes
Across diseases and different levels of ability	A set of publicly available, efficient, and flexible measurements of PROs, including health-related quality of life for the clinical research community	Doubtful	Yes

did not contain any information on PROM development, two were reviews, two were methodological articles, two were written in languages not readable by the review team (i.e., Korean and Chinese), two were on PROMs not included in this review, one was not retrievable, and one linked the content of a PROM to a health framework without focusing on physical functioning. The 112 included records comprised 22 focusing on PROM development [18,33–53], while the remaining records included 62 studies on content validity [54–115], and 41 studies on structural validity [60–62,68,70,71,82,90,95–98,107,116–143].

3.1. Content validity

3.1.1. Quality of PROM development studies

Table 2 presents a summary of the development studies describing construct definition, target population, and intended context of use of the 17 PROMs. Nine studies were specifically developed to measure physical functioning for LBP (Table 2). Concept elicitation was deemed inadequate for 10 PROMs because no patients were involved in their development (Table 2). For the other PROMs (i.e., BPI-PI, QBPDS, and the five PROMIS-PF short forms), it was

doubtful because the included patients were most likely not representative of the target population. For BPI-PI, the number and patients' characteristics included in the development were not reported; for QBPDS, 34 patients with LBP attending an orthopedic clinic or a rehabilitation center were interviewed, but their characteristics were not reported; for the PROMIS-PF item bank, 15 patients were involved in focus groups and evaluation: 8 with self-reported arthritis, 7 with unclear diagnosis, 12 female, 14 Caucasian, age range 31–80 years, all with at least high school education.

Only the development of the QBPDS and PROMIS-PF featured cognitive interviews with patients. The QBPDS involved 242 patients with LBP, but the process was judged to be inadequate because the questionnaire was not tested in its final form. Cognitive interviews were of doubtful quality for PROMIS-PF because the patients included were most likely not representative of the target population (i.e., 18 patients with unclear diagnosis: 12 female, 12 Caucasian, age range 48–93 years, 12 with at least high school education).

3.1.2. Quality and results of content validity studies

The 62 articles on content validity comprised 76 studies, 69 involving patients and seven involving professionals (Appendix C). No studies on content validity of RMDQ-18, MPI-PI, and SF36-PF, and the five PROMIS-PF short forms were found. Three studies involving patients (4%), all of adequate quality, truly aimed to assess (multiple aspects of) the content validity: two assessed relevance [66,72], one comprehensibility [66], and one comprehensiveness of RMDQ-24 [72]; another study relevance and comprehensibility of BPI-PI [99] (Table 3). All other (96%) content validity studies involving patients were cross-cultural adaptations that included a pretest of the translated PROMs: most on ODI 2.1a and RMDQ-24, several on QBPDS, and one each on ODI 1.0, CLBPDQ, MLBPDQ, RMDQ-23, and LBPRS-DI (Appendix C). All these studies were of doubtful quality, whereas the other 12 studies could not be rated because it was unclear which aspect was assessed. Fifty-four studies assessed comprehensibility within a cross-cultural adaptation; two studies (studies by Alnahhal and May [55] and Christakou et al. [61]) assessed content relevance of QBPDS, and the former ([55]) also assessed its comprehensiveness. The reporting of results of the pretest phase of cross-cultural adaptations was very limited, especially when compared to the reporting of the three studies specifically designed to evaluate content validity [66,72,99] (Appendix C).

The seven content validity studies involving professionals were also embedded within a cross-cultural adaptation study. All studies were of doubtful quality, and results were poorly reported. Five assessed relevance (ODI 1.0, ODI 2.1a, RMDQ-24, QBPDS) [58,60,61,101], and the remaining two assessed comprehensiveness (ODI 1.0 and RMDQ-24) [58] (Appendix D).

3.1.3. Evidence synthesis

High-quality evidence was only available for the RMDQ-24. It displayed insufficient results for comprehensiveness (based on one adequate quality study [72] and reviewers' rating) and sufficient results for

Table 3. Evidence synthesis on the content and structural validity of PROMs to measure physical functioning in patients with low back pain

	Content validity						Structural validity	
	Relevance		Comprehensiveness		Comprehensibility			
PROM	Rating of results	Quality of evidence	Rating of results	Quality of evidence	Rating of results	Quality of evidence	Rating of results	Quality of evidence
RMDQ-24	+	Very low	−	High	+	High	−	High
RMDQ-23	+	Very low	−	Very low	+	Very low	±	Moderate
RMDQ-18	+	Very low	−	Very low	+	Very low	−	High
ODI 1.0	±	Very low	−	Very low	+	Very low	−	High
ODI 2.1a	±	Very low	−	Very low	+	Very low	±	Moderate
CLBPDQ	±	Very low	−	Very low	+	Very low	?	?
MLBPDQ	±	Very low	−	Very low	+	Very low	?	?
BPI-PI	±	Very low	−	Very low	+	Very low	+	Moderate
MPI-PI	±	Very low	−	Very low	+	Very low	±	Moderate
PF-SF36	+	Very low	−	Very low	+	Very low	−	Moderate
LBPRS-DI	±	Very low	−	Very low	+	Very low	?	?
QBPDS	+	Low	+	Very low	+	Very low	±	Moderate
PROMIS-PF-4	+	Low	−	Very low	+	Low	?	?
PROMIS-PF-6	+	Low	−	Very low	+	Low	?	?
PROMIS-PF-8	+	Low	−	Very low	+	Low	?	?
PROMIS-PF-10	+	Low	+	Very low	+	Low	?	?
PROMIS-PF-20	+	Low	+	Very low	+	Low	?	?

Abbreviations: PROMs, patient-reported outcome measures; +, satisfactory results; −, unsatisfactory results; ±, inconsistent results; ?, indeterminate.

comprehensibility (based on one adequate quality study [66] and reviewers' rating); RMDQ-24 relevance rating was supported by very low quality evidence (Table 3). For all other PROMs, ratings were underpinned by (very) low-quality evidence (Table 3). Evidence quality resulted "very low" when development and content validity studies provided indeterminate ratings because only based on reviewers' rating (Terwee et al., 2017, unpublished data). For QBPDS and PROMIS-PF, short forms of the quality of evidence for relevance was "low" based on sufficient results in reviewers' ratings of development and content validity studies. Quality of evidence on comprehensiveness was rated as "very low" for all the instruments (except RMDQ-24, Table 3). For all PROMs, comprehensibility was rated as sufficient, but quality of evidence was very low: despite the large amount of studies for some PROMs (i.e., ODI 2.1a and QBPDS), reporting of results was very limited, and this led to indeterminate ratings in these studies. Based on all these results (Table 3), it is not possible to establish which PROM has the best content validity for physical functioning in nsLBP patients.

3.2. Structural validity

3.2.1. Quality and results of studies

Forty-six studies assessed structural validity of the PROMs in LBP patients: most assessed ODI 2.1a or RMDQ-24, with four to zero studies for each of the other instruments (Table 4). Most authors used exploratory factor analysis (EFA, 48%) or an IRT Rasch model (41%); the remainder used confirmatory factor analysis or an IRT 2-parameter logistic model (Table 4). Most of the studies (59%) were of at least adequate quality (26% doubtful and 15% inadequate). Eight study results could not be scored on the unidimensionality criteria because of incomplete reporting of EFA eigenvalues or explained variance [71,82,96,97,107,116,137,139], of a priori determined parameters of Rasch analysis (Appendix B) [118–120], or of CFA results because it was applied to the whole scale and not to the subscale of interest for this review (MPI-PI) [117].

3.2.2. Evidence synthesis

All studies on RMDQ-24, RMDQ-18, and ODI 1.0 displayed negative results, providing high-quality evidence of insufficient unidimensionality (Table 3). Studies of at least adequate quality for RMDQ-23, ODI 2.1a, MPI-PI, and QBPDS [60,61,70,90,95,98,121,129,134,138] showed inconsistent results, indicating uncertain unidimensionality underpinned by moderate quality evidence. BPI-PI proved to be the only PROM with sufficient unidimensionality (moderate quality evidence), whereas unidimensionality for SF36-PF was insufficient (moderate quality, Table 3). All the other PROMs displayed indeterminate ratings (Table 3).

4. Discussion

To measure physical functioning in patients with LBP, the RMDQ-24 displayed high-quality evidence for sufficient comprehensibility and for insufficient comprehensiveness, whereas the evidence quality of its relevance is very low. Low to very low quality evidence was found on the content validity of the other 16 PROMs included in this review, suggesting that current knowledge on this measurement property is very uncertain. The BPI-PI displayed moderate quality evidence for sufficient unidimensionality; conflicting findings based on moderate quality evidence were found for RMDQ-23, ODI 2.1a, MPI-PI, and QBPDS, and insufficient findings on high- to moderate-quality evidence for RMDQ-24, RMDQ-18, ODI 1.0, and SF36-PF.

The inadequate evidence quality underpinning the content validity of many routinely used and/or recommended PROMs to measure physical functioning in patients with LBP is worrisome, as it is probably the first measurement property to consider when selecting a PROM [14]. We found only three studies specifically aimed at assessment of this measurement property of PROMs in LBP [66,72,99]. Fortunately, these three qualitative studies were of adequate quality and can serve as example for future research. In addition, for future content validity studies, we suggest to consult the recently developed COSMIN standards on this measurement property (Terwee et al., 2017, unpublished data). A previous systematic review on the measurement properties of RMDQ-24 and ODI 2.1a documented the lack of head-to-head comparisons on content validity of these PROMs in the same LBP patients [21], and this unmet need should also be filled by future qualitative research.

The RMDQ-24 is the most frequently used tool to assess physical functioning in patients with LBP [6], but this review raises concerns on its content validity (specifically comprehensiveness) and its unidimensionality. Hush et al. [72] found that it may be an instrument more appropriate in acute LBP as many items refer to self-care activities, whereas leisure and exercise activities (potentially more relevant for patient with chronic LBP) are not captured; this omission may impact RMDQ-24's responsiveness. However, the RMDQ-24 scored well on comprehensibility and other measurement properties [21], and its relevance (arguably the most important aspect of content validity) for physical functioning in patients with LBP has not been adequately explored yet. Our conclusion on its lack of unidimensionality contrasts with prior studies included in this review concluding that it is a unidimensional tool [127,131,143], but these studies presented some issues. First, their conclusions were drawn despite findings highlighting misfit to the Rasch model for some items. Second, none adequately compared the model fit of alternative factor structures with the same statistical technique (e.g., CFA). In light of these considerations, although it is apparent that the RMDQ-24 is not fully unidimensional, it remains unclear what is the influence of this deviation on the total score routine use.

Table 4. Characteristics, quality assessment, and results of the structural validity studies in patients with low back pain

			Patients' characteristics						COSMIN quality rating		
PROM	Reference	Country (language)	*n*	Female (%)	Age (yr, μ±SD)	Pain duration	Work status	PROM score (μ ± SD)		Analysis - model	Results (synthesis)
ODI 1.0	Fisher, 1997 [126]	United Kingdom (English)	190	63	43 ± 12	8 ± 7 yr			Very good	PCA with VR on factors with eigenvalues >1	Two factors accounted for 33% and 12% of variance, respectively.
	Page, 2002 [136]	United States (English)	95	55	48 ± 16	2 ± 1 wk		33 ± 18	Inadequate	Rasch (model not specified)	Item 1, infit mean square statistic = 1.58, outfit mean square statistic = 1.66. Fit statistics were not presented for other items but assumable not ≤0.5 or ≥1.5.
	White, 2002 [141]	United States (English)	942	47	41 ± 14	21% <1 mo 49% 1–6 mo 30% >6 mo	47% empl 35% offwork 18% unempl		Very good	Rasch PCM	Item 1, infit mean square statistic = 1.5, outfit mean square statistic = 1.6. Fit statistics not presented for other items but assumable not ≤0.5 or ≥1.5. Disordered threshold for category 1 of item 6.
	Lochhead, 2013 [132]	Canada (English)	133	42			100% offwork		Adequate	Rasch PCM	No items displayed infit or outfit mean square statistics ≤0.5 or ≥1.5. Z-standardized values of five items (1, 2, 5, 7, and 8) were ≤−2 or ≥2. Disordered thresholds for: category 1 (items 5, 7, 9), category 2 (1, 4, 8).
ODI 2.1a	Chow, 2005 [60]	Hong Kong (Chinese)	75	33	42 ± 10	14 ± 10 mo		46 ± 16	Adequate	PCA with VR	Two factors accounted for 49% and 11% of the variance, respectively.
	Guermazi, 2005 [70]	Arabic (Tunisia)[a]	80	54	47 ± 13	8 ± 5 mo	41% empl	39 ± 16	Adequate	PCA with VR on factors with eigenvalues >1	Two factors accounted for 32% and 27% of the variance, respectively.
	Osthus, 2006 [95]	German (Germany)	160	27	48 ± 9				Adequate	PCA	Two factors each accounted for 30% of the variance.
	Davidson, 2008 [122]	Australia (English)	100	64	53 ± 15	31% <6 wk 17% <3 mo 52% >3 mo	33% empl 5% sick 10% disab 50% unempl 2% stud		Inadequate	Rasch PCM	Good item fit for all items (χ^2: $P > 0.01$). None had fit residuals ≤2.5 or ≥2.5. Good overall fit (χ^2 item-trait interaction: $P = 0.14$). Disordered

											thresholds for: category 1 (items 2 and 9), category 2 (6, 8), category 3 (8), and category 4 (9). Suboptimal targeting.
	Monticone, 2009 [90]	Italian (Italy)	126	58	47 ± 14	21 ± 20 mo	82% empl	?	Adequate	EFA extracting factors with eigenvalues >1	One factor accounted for 45% of the variance.
	Klemenc-Ketis, 2011 [130]	Slovene (Slovenia)	129	47	50 ± 10	116 ± 110 mo	71% empl 27% retired 2% unempl	31 ± 17	Doubtful	PCA	One factor accounted for 49% of the variance.
	Payares, 2011 [137]	Spanish (Colombia)	111	68	45 ± 16				Doubtful	EFA with VR	Two factors accounted together for 56% of the variance.
	Pekkanen, 2011 [97]	Finnish (Finland)	115	44	49 ± 13	27% <6 mo 39% <24 mo 34% >24 mo			Doubtful	EFA with PR	Two factors accounted together for 51% of the variance.
	Algarni, 2014 [116]	Saudi Arabia (Arabic)	100	45	40 ± 13		12% unempl		Doubtful	EFA with VR	Two factors accounted together for 63% of the variance.
	Van Hooff, 2015 [107]	Dutch (Nether-lands)	244	57	46 ± 11	13 ± 10	79% empl 21% unempl	40 ± 12	Adequate	EFA with VR	One factor accounted for 36% of the variance.
	Gamus, 2016 [68]	Hebrew (Israel)	115	56	51 ± 16	100% >6 wk		18 ± 9	Doubtful	EFA	One factor accounted for 65% of the variance.
	Brodke, 2016 [118]	English (United States)	1,610	53	57 ± 17			39 ± 19	Doubtful	Rasch (model not specified)	No items with outfit statistic ≤0.5 or ≥1.5, but infit statistics were not reported. Inadequate targeting.
	Brodke, 2017 [119]	English (United States)	1,607	53	57 ± 17			39 ± 19	Doubtful	Rasch (model not specified)	No items with outfits statistic ≤0.5 or ≥1.5, but infit statistics were not reported. Inadequate targeting.
MLBPDQ	Choi, 2015 [120]	English (United States)	42	69	54 ± 20	33% <1 yr 12% 1-4 yr 48% >4 yr			Inadequate	Rasch (model not specified)	
RMDQ-24	Exner, 2000 [125]	German (Switzer-land)	282						Adequate	PCA with VR on factors with eigenvalues >1	Six factors were extracted, each accounting between 6% and 10% of the variance.
	Kucukdeveci, 2001 [131]	Turkish (Turkey)	81	63	37 ± 11	5 ± 4 yr		15	Doubtful	Rasch model for dichotomous data	Item 10, outfit mean square statistic = 1.76. Item 15, outfit mean square statistic = 0.46. Item 19, outfit mean square statistic = 0.29. Infit and outfit statistics for the other items were not ≤0.5 or ≥1.5.

(*Continued*)

Table 4. Continued

PROM	Reference	Country (language)	Patients' characteristics						COSMIN quality rating	Analysis - model	Results (synthesis)
			n	Female (%)	Age (yr, μ±SD)	Pain duration	Work status	PROM score (μ ± SD)			
	Williams, 2001 [142]	English (Canada)	94	27	37 ± 11	6 ± 11 wk			Inadequate	EFA	First factor accounted for 27% of the variance, second 9%, third 10% and fourth 9%.
	Garratt, 2003 [127]	English (United Kingdom)	1008	55	43			9 ± 4	Adequate	Rasch model for dichotomous data	Item 2, outfit mean square statistic = 1.68. Infit and outfit statistics for the other items were not ≤0.5 or ≥1.5.
	Suzukamo, 2003 [139]	Japanese (Japan)	214	46	53 (21–86)			9 ± 5	Adequate	PCA	Several factors were extracted with eigenvalue >1. Explained variance was not reported.
	Mâaroufi, 2007 [82]	Moroccan (Morocco)	76	91	51 ± 10	62 mo		15 ± 5	Inadequate	Multiple correspondence analysis	One factor accounted for 22% of the variance. Other four factors accounted together for 28% of the variance.
	Davidson, 2009 [123]	English (Australia)	140	66	51 ± 17	43% >6 wk	41% empl		Adequate	Rasch model for dichotomous data	Good item fit for all items (χ^2: $P > 0.01$). Item 9 presented a fit residuals = −2.5. Good overall fit (χ^2 item-trait interaction: $P = 0.03$). Three item pairs were locally dependent (5-7, 5-23 and 7-12). Inadequate targeting.
	Grotle, 2013 [128]	Norwegian (Norway)	250	48	49 ± 11		74% empl	9 ± 4	Very good	Rasch model for dichotomous data	Poor fit for items 13 and 23 (χ^2: $P < 0.01$). Four items displayed fit residuals ≤−2.5 or ≥2.5 (items 3, 13, 18, 23). Poor overall fit (χ^2 item-trait interaction: $P < 0.001$). Adequate targeting.
	Nambi, 2013 [135]	Guajarati (India)	30	67	46 ± 12			19	Inadequate	Rasch model for dichotomous data	Item 7, outfit mean square statistic = 2.13. Item 14, outfit mean square

											statistic = 0.49. Item 19, outfit mean square statistic = 1.79. Item infit and outfit statistics for the other items were not ≤0.5 or ≥1.5.
	Guic, 2014 [71]	Spanish (Chile)	206	25	37 (18-64)				Adequate		One factor accounted for 20% of the variance.
	Magnussen, 2015[b] [133]	Norwegian (Norway)	371	47	39 ± 11			8 ± 4	Adequate	EFA with OFR	A 1-factor solution had poor fit (CFI = 0.81, RMSEA = 0.07). A 3-factor solution gave the best fit (CFI = 0.96, RMSEA = 0.05).
	Payares, 2015 [137]	Spanish (Colombia)	133	68		37% acute 30% subacute 33% chronic		9 ± 5	Doubtful	EFA with VR	
	Yamato, 2017 [143]	English (Australia)	2826	50	46 ± 16			12 ± 5	Very good	Rasch model for dichotomous data	Item 2, outfit mean square statistic = 1.7. Infit and outfit statistics for the other items were not ≤0.5 or ≥1.5. No locally dependent item pairs. Disordered ordering for item 2, but unclear if assessed for every item. Adequate targeting.
RMDQ-23	Cook, 2008 [121]	English (United States)	874	53	47 ± 13		46% empl 20% retired 4% unempl 18% work compens		Very good	CFA and EFA	CFA: 1-factor solution gave the following fit: CFI = 0.90, RMSEA = 0.08, SRMR = 0.09. EFA: One factor accounted for 51% of the variance.
									Adequate	2-PLM for dichotomous data	Good model fit (S-X2: $P < 0.01$; S-G2: $P < 0.01$).
	Kent, 2015 [129]	English (United Kingdom)	500	56	46 (IQR: 39-53)	8 mo (IQR: 4–15)			Adequate	Rasch model for dichotomous data	Item 23 poor fit (χ^2: $P < 0.01$). Three items (3, 18 and 22) displayed fit residuals ≤ −2.5 or ≥ 2.5. Poor overall fit (χ^2 item-trait interaction: $P < 0.001$). Three locally dependent item pairs (4-22, 5-18 and 7-13). Suboptimal targeting.

(*Continued*)

Table 4. Continued

PROM	Reference	Country (language)	Patients' characteristics						COSMIN quality rating	Analysis - model	Results (synthesis)
			n	Female (%)	Age (yr, μ±SD)	Pain duration	Work status	PROM score (μ ± SD)			
		Danish (Denmark)	500	54		42% < 3 mo			Adequate	Rasch model for dichotomous data	Poor fit for four items (14, 15, 21, 23; χ^2: $P < 0.01$). Six items (3, 7, 14, 17, 21 and 23) displayed fit residuals ≤ -2.5 or ≥ 2.5. Poor overall fit (χ^2 item-trait interaction: $P < 0.001$). Four locally dependent item pairs (3-18, 6-10, 7-13, and 6-22). Inadequate targeting.
	Mielenz, 2015 [134]	English (United States)	670						Very good	EFA with OQR and CFA	EFA: one factor with eigenvalue >4/1 ratio with other eigenvalues. CFA: good fit 1-factor solution (RMSEA = 0.058, CFI = 0.99, GFI = 0.987, SRMR = 0.075).
RMDQ-18	Davidson, 2009[c] [123]	English (Australia)	140	66	51 ± 17	43% > 6 wk	41% empl		Doubtful Adequate	2-PLM Rasch model for dichotomous data	Item 9 displayed poor fit (χ^2: $P > 0.01$). Item 9 presented a fit residuals = −2.5. Good overall fit (χ^2: $P = 0.03$). Three locally dependent item pairs (5-7, 5-23, and 7-12). Inadequate targeting.
RMDQ-18	Grotle, 2013[c] [128]	Norwegian (Norway)	250	48	48 ± 11		74 empl	9 ± 4	Very good	Rasch model for dichotomous data	Items 13 and 23 displayed poor fit (χ^2: $P > 0.01$). Four items (3, 13, 18 and 23) displayed fit residuals ≤ -2.5 or ≥ 2.5. Poor overall fit (χ^2: $P < 0.001$). Inadequate targeting.
BPI-PI	Tan, 2004 [140]	English (United States)	444					8 ± 2	Adequate	PCA with PR	PCA was applied to all BPI items. Two factors accounted together for 68% of the total

											variance. The first factor accounted for 51%, the second for 13%.
MPI-PI	Bernstein, 1995 [117]	English (United States)	194	52					Inadequate	CFA with oblique multiple groups method	CFA was applied to section 1 of the MPI. Five factors accounted for 68% of the variance.
	Riley, 1999 [138]	English (United States)	472	39	44 ± 14	3 ± 4 yr			Very good	EFA with OR	EFA was applied to section 1 of the MPI. Four factors accounted for 70% of the variance. First factor accounted for 41% of the variance, second 14%, third 8%, and fourth 6%.
			346	56	44 ± 11	5 ± 5 yr			Adequate	CFA	A 4-factor solution for section 1 of the MPI was tested. This solution gave the following fit: $\chi^2 = 491.22$ ($P < 0.001$), ECVI = 1.66, GFI = 0.97, NFI = 0.96, PNFI = 0.80, and RMSEA = 0.98.
SF36-PF	Davidson, 2004 [124]	English (Australia)	140	66	51 ± 17	11% <1 wk 32% 1–6 wk 22% 6wk–6 mo 34% >6 mo	41% empl 10% unempl 49% offwork		Adequate	Rasch PCM	Item 1, outfit mean square statistic = 2.17. Item 10, infit mean square statistic = 1.51, outfit mean square statistic = 2.16. Infit and outfit statistics for the other items were not ≤0.5 or ≥1.5.
	Brodke, 2017 [119]	English (United States)	1607	53	57 ± 17			40 ± 28	Doubtful	Rasch (model not specified)	Item 1, outfits statistic > 1.5, other items no outfit statistics ≤ 0.5 or ≥ 1.5. Infit statistics not reported. Inadequate targeting.
QBPDS	Christakou, 2011 [61]	Greek (Greece)	130	54	41 ± 12	39 ± 37 mo	78% empl 5% retired 17% housewifes		Adequate	EFA	Six factors accounted together for 82% of the variance. The first accounted for 16% of the variance, second 30%, third 12%, fourth 11%, fifth 7%, and sixth 6%.
	Cruz, 2013 [62]	Portuguese (Portugal)	132	73	47 ± 13	14% 3–6 mo 20% 6–24 mo 66% >24 mo	66% empl 11% unempl 12% retired	36 ± 18	Doubtful	PCA	Four factors accounted together for 70% of the variance. First factor

(*Continued*)

Table 4. Continued

PROM	Reference	Country (language)	Patients' characteristics						COSMIN quality rating	Analysis - model	Results (synthesis)
			n	Female (%)	Age (yr, μ±SD)	Pain duration	Work status	PROM score (μ ± SD)			
											accounted for 52% of the variance, second 7%, third 5%, and fourth 5%.
	Riecke, 2016 [98]	German (Germany)	180	71		100% > 3 m			Adequate	EFA with OR	Four factors were accounted together for 57% of the variance. First factor accounted for 44%, second 6%, third 4%, and fourth 3%.

Abbreviations: PCA, principal component analysis; VR, varimax rotation; PCM, partial credit model; EFA, exploratory factor analysis; PR, promax rotation; OFR, oblique factor rotation; CFI, comparative fit index; RMSEA, root mean square error of approximation; CFA, confirmatory factor analysis; SRMR, standardized root mean residuals; 2-PLM, 2-parameter logistic model; IQR, interquartile range; OQR, oblique quartimax rotation; GFI, goodness-of-fit index; OR, oblimin rotation; ECVI, expected cross-validation index; NFI, normed fit index; PNFI, parsimony normed fit index.

Empty cells indicate data not available.

[a] Items 8 and 10 of ODI 2.1a were excluded from this Arabic version because they were very poorly endorsed.

[b] Items 19 and 24 were excluded from the analyses due to 0.8% endorsement of the 'yes' response option.

[c] Items number refers to the number in the 24-item version of the Roland Morris Disability Questionnaire.

This review also did not confirm the unidimensionality of the total score of other well-known physical functioning PROMs, i.e., RMDQ-23, ODI 2.1a, and QBPDS (Table 3) [5,6]. Psychometric studies with a large sample and framed within a multi-national consortium are needed, also to explore whether dimensionality varies in different cultures and populations. We support the suggestion made by others to apply bifactor analysis to assess the degree to which a PROM departs from unidimensionality [144–147]. In a bifactor model, items are allowed to load on both a general (unidimensional) factor and on group factors (i.e., potential subscales); the factor loadings obtained for the general and group factors can be used to estimate how much variance in the data is explained by the general factor and the relative impact of the group factors [144–147]. None of the studies included in this review adopted this statistical technique, or a multidimensional IRT model [148–150].

The NIH Task Force on research standards for chronic LBP recommended the PROMIS-PF-4 to measure physical functioning [22]. However, as the PROMIS-PF short forms are relatively new instruments in the LBP research arena, there is limited evidence on content validity, and no studies are available on structural validity in patients with LBP. Nevertheless, these instruments have exhibited sufficient measurement properties in the general population and in other musculoskeletal disorders [151–155]. Therefore, they can also be considered for use in LBP. Some authors have recently suggested the use of one among ODI 2.1a, QBPDS, and RMDQ-24 for physical functioning in patients with LBP [8]. These suggestions were mainly based on the results of an international, multidisciplinary, and multistakeholder Delphi survey [156] that was informed by the results of this and other systematic reviews [20,21] on measurement properties of physical functioning PROMs in LBP patients.

In this systematic review, we have specifically talked about "the validity of the PROMs". However, the reader should bear in mind that the validity of an instrument mainly concerns the interpretation of the instrument scores in a given application [157]; therefore, the results of this review may not be generalizable to every context. In addition, it should be noted that current perspectives on validity in other fields specifically focus only on the inferences, claims, or decisions made, based on an instrument scores and not the instrument itself [158]. It could be expected that LBP-specific PROMs such as RMDQ-24, ODI 2.1a, or QBPDS display better content validity than generic PROMs to measure physical functioning in patients with LBP; nevertheless, this review does not provide any evidence to support this expectation (Table 3). To resolve the debate on generic- vs. disease-specific instruments in this context, head-to-head content validity comparisons are needed in patients with LBP.

This is the first review to apply a novel method of assessment recently developed by the COSMIN initiative

(Terwee et al., 2017, unpublished data) to thoroughly assess the content validity of a set of PROMs, by taking into account the following: methodological quality and results of their development process, methodological quality and results of new content validity studies, and their content. Also, we decided to expand the evidence base with cross-cultural adaptation studies that included a comprehensibility assessment in their pilot test phase. This assessment has been recommended for PROMs cross-cultural adaptations to ensure that patients completing a questionnaire are able to understand its content as intended. Cross-cultural adaptation studies should continue to focus on published guidance [29], but authors could consider performing a more detailed assessment of this aspect, following specific guidance on how to assess comprehensibility of a PROM [15,16] (Terwee et al., 2017, unpublished data).

A potential weakness of our study is our decision to consider different language versions of the PROMs as the same questionnaire in the evidence syntheses. Our rationale was that there is usually no enough evidence to make a synthesis per language version because there is no evidence clearly indicating what is the most appropriate way to synthesize the evidence on PROMs measurement properties. However, for more detailed scrutiny, we provide the methodological quality and results of each study specifying language and country (Appendices C and D, Table 4). Another potential limitation is our focus on a limited set of 17 PROMs, chosen among those most frequently used and/or recommended for LBP [7−11,22]. Nevertheless, it would be very surprising to find moderate or high quality evidence on the validity of other instruments, but this cannot be ruled out completely. A limitation of this review is the impossibility to check the eligibility of two articles in languages inaccessible for the review team (i.e., Chinese and Korean).

Content validity is the first measurement property to consider when selecting a PROM [14], and this systematic review showed it is underinvestigated for PROMs to measure physical functioning in patients with LBP. Available evidence showed potentially important limitations in the unidimensionality of several widely used PROMs in patients with LBP, such as ODI 2.1a, RMDQ-24, RMDQ-23, SF36-PF, and QBPDS [5,6]. Future research should be dedicated to fill existing research gaps on content and structural validity, possibly by performing head-to-head comparison studies of more instruments measuring the same domain.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2017.11.005.

References

[1] Dieleman JL, Baral R, Birger M, Bui AL, Bulchis A, Chapin A, et al. US spending on personal health care and public health, 1996-2013. JAMA 2016;316:2627−46.

[2] Vos T, Allen C, Arora M, Barber RM, Bhutta ZA, Brown A, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet 2016;388:1545−602.

[3] Black N. Patient reported outcome measures could help transform healthcare. BMJ 2013;346:f167.

[4] Chiarotto A, Deyo RA, Terwee CB, Boers M, Buchbinder R, Corbin TP, et al. Core outcome domains for clinical trials in non-specific low back pain. Eur Spine J 2015;24:1127−42.

[5] Chapman JR, Norvell DC, Hermsmeyer JT, Bransford RJ, DeVine J, McGirt MJ, et al. Evaluating common outcomes for measuring treatment success for chronic low back pain. Spine 2011;36:S54−68.

[6] Froud R, Patel S, Rajendran D, Bright P, Bjørkli T, Buchbinder R, et al. A systematic review of outcome measures use, analytical approaches, reporting methods, and publication volume by year in low back pain trials published between 1980 and 2012: respice, adspice, et prospice. PLoS One 2016;11(10):e0164573.

[7] Bombardier C. Outcome assessments in the evaluation of treatment of spinal disorders: summary and general recommendations. Spine 2000;25:3100−3.

[8] Chiarotto A, Terwee CB, Ostelo RW. Choosing the right outcome measurement instruments for low back pain. Best Pract Res Clin Rheumatol 2016;30:1003−20.

[9] Clement RC, Welander A, Stowell C, Cha TD, Chen JL, Davies M, et al. A proposed set of metrics for standardized outcome reporting in the management of low back pain. Acta Orthopaedica 2015;86:523−33.

[10] Deyo RA, Battie M, Beurskens A, Bombardier C, Croft P, Koes B, et al. Outcome measures for low back pain research: a proposal for standardized use. Spine 1998;23:2003−13.

[11] Dworkin RH, Turk DC, Farrar JT, Haythornthwaite JA, Jensen MP, Katz NP, et al. Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. Pain 2005;113(1-2):9−19.

[12] Messick S. Test validity and the ethics of assessment. Am Psychol 1980;35:1012.

[13] Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol 2010;63:737−45.

[14] Prinsen CA, Vohra S, Rose MR, Boers M, Tugwell P, Clarke M, et al. How to select outcome measurement instruments for outcomes included in a "Core Outcome Set"−a practical guideline. Trials 2016;17:449.

[15] Brod M, Tesler LE, Christensen TL. Qualitative research and content validity: developing best practices based on science and experience. Qual Life Res 2009;18:1263.

[16] Patrick DL, Burke LB, Gwaltney CJ, Leidy NK, Martin ML, Molsen E, et al. Content validity—establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO Good Research Practices Task Force report: part 2—assessing respondent understanding. Value Health 2011;14:978−88.

[17] Patrick DL, Burke LB, Gwaltney CJ, Leidy NK, Martin ML, Molsen E, et al. Content validity—establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO good research practices task force report: part 1—eliciting concepts for a new PRO instrument. Value Health 2011;14:967−77.

[18] Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005−2008. J Clin Epidemiol 2010;63:1179−94.

[19] Gerbing DW, Anderson JC. An updated paradigm for scale development incorporating unidimensionality and its assessment. J Marketing Res 1988;25:186−92.
[20] Speksnijder CM, Koppenaal T, Knottnerus JA, Spigt M, Staal JB, Terwee CB. Measurement properties of the quebec back pain disability scale in patients with Nonspecific low back pain: systematic review. Phys Ther 2016;96:1816−31.
[21] Chiarotto A, Maxwell LJ, Terwee CB, Wells GA, Tugwell P, Ostelo RW. Roland-Morris Disability Questionnaire and Oswestry Disability Index: which has better measurement properties for measuring physical functioning in nonspecific low back Pain? Systematic review and meta-analysis. Phys Ther 2016; 96:1620−37.
[22] Deyo RA, Dworkin SF, Amtmann D, Andersson G, Borenstein D, Carragee E, et al. Report of the NIH Task Force on research standards for chronic low back pain. Pain Med 2014;15:1249−67.
[23] Boers M, Kirwan JR, Wells G, Beaton D, Gossec L, d'Agostino M-A, et al. Developing core outcome measurement sets for clinical trials: OMERACT filter 2.0. J Clin Epidemiol 2014;67:745−53.
[24] Williamson PR, Altman DG, Blazeby JM, Clarke M, Devane D, Gargon E, et al. Developing core outcome sets for clinical trials: issues to consider. Trials 2012;13:132.
[25] Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Plos Med 2009;6:e1000097.
[26] Kamper SJ, Apeldoorn AT, Chiarotto A, Smeets RJ, Ostelo RW, Guzman J, et al. Multidisciplinary biopsychosocial rehabilitation for chronic low back pain. Cochrane Database Syst Rev 2014; CD000963.
[27] Saragiotto BT, Maher CG, Yamato TP, Costa LO, Menezes Costa LC, Ostelo RW, et al. Motor control exercise for chronic non-specific low-back pain. Cochrane Database Syst Rev 2016;CD012004.
[28] Terwee CB, Jansma EP, Riphagen II, de Vet HC. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. Qual Life Res 2009;18:1115−23.
[29] Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine 2000;25:3186−91.
[30] Fairbank JC. Why are there different versions of the Oswestry Disability Index? A review. J Neurosurg Spine 2014;20(1):83−6.
[31] Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res 2012;21:651−7.
[32] Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336:924.
[33] PROMIS instrument development and validation scientific standards version 2.0;2013:1−72. Available at http://www.healthmeasures.net/images/PROMIS/PROMISStandards_Vers2.0_Final.pdf.
[34] Baker D, Pynsent P, Fairbank J. The oswestry disability index revisited: its reliability, repeatability and validity, and a comparison with the St Thomas's disability index. In: Roland M, Jenner J, editors. Back pain: new approaches to rehabilitation and education. Manchester (UK): Manchester University Press; 1989:174−86.
[35] Bruce B, Fries JF, Ambrosini D, Lingala B, Gandek B, Rose M, et al. Better assessment of physical function: item improvement is neglected but essential. Arthritis Res Ther 2009;11(6):R191.
[36] Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. Med Care 2007;45:S3.
[37] Cleeland CS. The Brief Pain Inventory. Available at https://www.mdanderson.org/documents/Departments-and-Divisions/Symptom-Research/BPI_UserGuide.pdf2009. Accessed December 20, 2017.
[38] Cleeland CS, Ryan K. Pain assessment: global used of the brief pain inventory. Ann Acad Med Singapore 1994;23:129−38.
[39] Daut RL, Cleeland CS, Flanery RC. Development of the Wisconsin Brief Pain Questionnaire to assess pain in cancer and other diseases. Pain 1983;17(2):197−210.
[40] DeWalt DA, Rothrock N, Yount S, Stone AA. Evaluation of item candidates: the PROMIS qualitative item review. Med Care 2007; 45:S12.
[41] Fairbank J, Couper J, Davies J, O'brien J. The Oswestry low back pain disability questionnaire. Physiotherapy 1980;66:271−3.
[42] Fritz JM, Irrgang JJ. A comparison of a modified Oswestry low back pain disability questionnaire and the Quebec back pain disability scale. Phys Ther 2001;81:776.
[43] Hudson-Cook N, Tomes-Nicholson K, Breen A. The revised Oswestry low-back pain disability questionnaire. In: Roland M, Jenner J, editors. Back pain: new approaches to rehabilitation and Education. New York: Manchester University Press; 1989: 187−204.
[44] Kerns RD, Turk DC, Rudy TE. The west haven-yale multidimensional pain inventory (WHYMPI). Pain 1985;23:345−56.
[45] Kopec JA, Esdaile JM, Abrahamowicz M, Abenhaim L, Wood-Dauphinee S, Lamping DL, et al. The Quebec back pain disability scale: conceptualization and development. J Clin Epidemiol 1996; 49:151−61.
[46] Manniche C, Asmussen K, Lauritsen B, Vinterberg H, Kreiner S, Jordan A. Low Back Pain Rating scale: validation of a tool for assessment of low back pain. Pain 1994;57:317−26.
[47] Meade T, Browne W, Mellows S, Townsend J, Webb J, North W, et al. Comparison of chiropractic and hospital outpatient management of low back pain: a feasibility study. J Epidemiol Community Health 1986;40:12−7.
[48] Patrick DL, Deyo RA, Atlas SJ, Singer DE, Chapin A, Keller RB. Assessing health-related quality of life in patients with sciatica. Spine 1995;20:1899−908.
[49] Roland M, Morris R. A study of the natural history of back pain: part I: development of a reliable and sensitive measure of disability in low-back pain. Spine 1983;8:141−4.
[50] Rose M, Bjorner JB, Gandek B, Bruce B, Fries JF, Ware JE. The PROMIS Physical Function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. J Clin Epidemiol 2014;67:516−26.
[51] Stewart A, Kamberg C. Physical functioning measures. In: Stewart A, Ware JE Jr, editors. Measuring functioning and well-being: the Medical Outcomes Study approach. Durham, NC: Duke University Press; 1992:86−101.
[52] Stratford PW, Binkley JM. Measurement properties of the RM-18: a modified version of the roland-morris disability scale. Spine 1997; 22:2416−21.
[53] Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36): I. Conceptual framework and item selection. Med Care 1992;30:473−83.
[54] Albert H, Jensen A, Dahl D, Rasmussen M. Criteria validation of the Roland Morris questionnaire. A Danish translation of the international scale for the assessment of functional level in patients with low back pain and sciatica. Ugeskr Laeger 2003; 165:1875−80.
[55] Alnahhal A, May S. Validation of the Arabic version of the Quebec back pain disability scale. Spine 2012;37:E1645−50.
[56] Baradaran A, Ebrahimzadeh MH, Birjandinejad A, Kachooei AR. Cross-cultural adaptation, validation, and reliability testing of the modified oswestry disability questionnaire in Persian population with low back pain. Asian Spine J 2016;10:215−9.
[57] Bendeddouche I, Rostom S, Bahiri R, Boudali A, Srifi N, Mawani N, et al. Translation, adaptation and validation of the Moroccan version of the quebec back pain disability scale. Clin Rheumatol 2012;31:943−9.

[58] Boscainos PJ, Sapkas G, Stilianessi E, Prouskas K, Papadakis SA. Greek versions of the oswestry and roland-morris disability questionnaires. Clin orthopaedics Relat Res 2003;411:40−53.
[59] Chhabra HS, Kapoor KS. New modified English and Hindi Oswestry Disability Index in low back pain patients treated conservatively in Indian population. Asian Spine J 2014;8:632−8.
[60] Chow JH, Chan CC. Validation of the Chinese version of the oswestry disability index. Work 2005;25:307−14.
[61] Christakou A, Andriopoulou M, Asimakopoulos P. Validity and reliability of the Greek version of the quebec back pain disability scale. J Back Musculoskelet Rehabil 2011;24:145−54.
[62] Cruz EB, Fernandes R, Carnide F, Vieira A, Moniz S, Nunes F. Cross-cultural adaptation and validation of the quebec back pain disability scale to european Portuguese language. Spine 2013;38: E1491−7.
[63] De Beer N, Stewart A, Becker P. Validation of the Tswana versions of the roland-morris disability questionnaire, quebec disability scale and Waddell disability index. South Afr J Physiother 2008;64: 23−30.
[64] Denis I, Fortin L. Development of a French-Canadian version of the Oswestry Disability Index: cross-cultural adaptation and validation. Spine 2012;37:E439−44.
[65] Fan S, Hu Z, Hong H, Zhao F. Cross-cultural adaptation and validation of simplified Chinese version of the Roland-Morris Disability Questionnaire. Spine 2012;37:875−80.
[66] Froud R, Ellard D, Patel S, Eldridge S, Underwood M. Primary outcome measure use in back pain trials may need radical reassessment. BMC Musculoskelet Disord 2015;16:88.
[67] Fujiwara A, Kobayashi N, Saiki K, Kitagawa T, Tamai K, Saotome K. Association of the Japanese Orthopaedic Association score with the Oswestry Disability Index, Roland-Morris Disability Questionnaire, and short-form 36. Spine 2003;28:1601−7.
[68] Gamus D, Glasser S, Langner E, Beth-Hakimian A, Caspi I, Carmel N, et al. Psychometric properties of the Hebrew version of the Oswestry Disability Index. J Back Musculoskelet Rehabil 2016;1−9.
[69] Grotle M, Brox J, Vollestad N. Cross-cultural adaptation of the Norwegian versions of the Roland-Morris Disability Questionnaire and the Oswestry Disability Index. J Rehabil Med 2003;35:241−7.
[70] Guermazi M, Mezghani M, Ghroubi S, Elleuch M, Poiraudeau S, Mrabet F, et al. The Oswestry index for low back pain translated into Arabic and validated in a Arab population. Ann Readapt Med Phys 2005;48:1−10.
[71] Guic E, Galdames S, Rebolledo P. Adaptación cultural y validación de la versión chilena del Cuestionario de Discapacidad Roland-Morris. Revista médica de Chile 2014;142:716−22.
[72] Hush JM, Refshauge KM, Sullivan G, De Souza L, McAuley JH. Do numerical rating scales and the Roland-Morris Disability Questionnaire capture changes that are meaningful to patients with persistent back pain? Clin Rehabil 2010;24(7):648−57.
[73] Jeon C-H, Kim D-J, Kim S-K, Kim D-J, Lee H-M, Park H-J. Validation in the cross-cultural adaptation of the Korean version of the oswestry disability index. J Korean Med Sci 2006;21:1092−7.
[74] Joshi VD, Raiturker PPP, Kulkarni AA. Validity and reliability of English and Marathi Oswestry Disability Index (version 2.1 a) in indian population. Spine 2013;38:E662−8.
[75] Kim D-Y, Lee S-H, Lee H-Y, Lee H-J, Chang S-B, Chung S-K, et al. Validation of the Korean version of the Oswestry Disability Index. Spine 2005;30:E123−7.
[76] Kim K-E, Lim J-Y. Cross-cultural adaptation and validation of the Korean version of the Roland-Morris Disability Questionnaire for use in low back pain. J Back Musculoskelet Rehabil 2011;24:83−8.
[77] Kovacs FM, Llobera J, del Real MTG, Abraira V, Gestoso M, Fernández C. Validation of the Spanish version of the Roland-Morris Questionnaire. Spine 2002;27:538−42.
[78] Lauridsen HH, Hartvigsen J, Manniche C, Korsholm L, Grunnet-Nilsson N. Danish version of the Oswestry Disability Index for patients with low back pain. Part 1: cross-cultural adaptation, reliability and validity in two different populations. Eur Spine J 2006;15:1705−16.
[79] Lee JS, Lee DH, Suh KT, Kim JI, Lim JM, Goh TS. Validation of the Korean version of the Roland−Morris Disability Questionnaire. Eur Spine J 2011;20:2115−9.
[80] Liu H, Tao H, Luo Z. Validation of the simplified Chinese version of the Oswestry Disability Index. Spine 2009;34:1211−6.
[81] Lue Y-J, Hsieh C-L, Huang M-H, Lin G-T, Lu Y-M. Development of a Chinese version of the Oswestry Disability Index version 2.1. Spine 2008;33:2354−60.
[82] Mâaroufi H, Benbouazza K, Faïk A, Bahiri R, Lazrak N, Abouqal R, et al. Translation, adaptation, and validation of the Moroccan version of the Roland Morris Disability Questionnaire. Spine 2007;32:1461−5.
[83] Maki D, Rajab E, Watson PJ, Critchley DJ. Cross-cultural translation, adaptation, and psychometric testing of the Roland-Morris Disability Questionnaire into modern standard Arabic. Spine 2014;39:E1537−44.
[84] Mannion A, Junge A, Fairbank J, Dvorak J, Grob D. Development of a German version of the Oswestry Disability Index. Part 1: cross-cultural adaptation, reliability, and validity. Eur Spine J 2006;15: 55−65.
[85] Mbada CE, Idowu OA, Ogunjimi OR, Ayanniyi O, Orimolade EA, Oladiran AB, et al. Cross-cultural adaptation, reliability and validity of the Yoruba version of the Roland Morris Disability Questionnaire. Spine 2017;42:497−503.
[86] Melikoglu MA, Kocabas H, Sezer I, Bilgilisoy M, Tuncer T. Validation of the Turkish version of the Quebec back pain disability scale for patients with low back pain. Spine 2009;34: E219−24.
[87] Misterska E, Jankowski R, Glowacki M. Quebec Back Pain Disability Scale, Low Back Outcome Score and revised Oswestry low back pain disability scale for patients with low back pain due to degenerative disc disease: evaluation of Polish versions. Spine 2011;36:E1722−9.
[88] Mohan V, Prashanth G, Meravanigi G, Rajagopalan N, Yerramshetty J. Adaptation of the Oswestry Disability Index to Kannada language and evaluation of its validity and reliability. Spine 2016;41:E674−80.
[89] Monteiro J, Faísca L, Nunes O, Hipólito J. Questionário de incapacidade de Roland Morris: adaptação e validação para a população portuguesa com lombalgia. Acta Médica Portuguesa 2010;23: 761−6.
[90] Monticone M, Baiardi P, Ferrari S, Foti C, Mugnai R, Pillastrini P, et al. Development of the Italian version of the Oswestry Disability Index (ODI-I): a cross-cultural adaptation, reliability, and validity study. Spine 2009;34:2090−5.
[91] Moon J, Kim YC, Park SY, Lee SC, Choi SP, Nahm FS, et al. Psychometric characteristics of the Korean version of the Roland-Morris Disability Questionnaire. J Korean Med Sci 2011;26: 1364−70.
[92] Mousavi SJ, Parnianpour M, Mehdian H, Montazeri A, Mobini B. The Oswestry Disability Index, the Roland-Morris Disability Questionnaire, and the Quebec Back Pain Disability Scale: translation and validation studies of the Iranian versions. Spine 2006;31: E454−9.
[93] Nuhr M, Crevenna R, Quittan M, Auterith A, Wiesinger G, Brockow T, et al. Cross-cultural adaption of the Manniche questionnaire for German-speaking low back pain patients. J Rehabil Med 2004;36:267−72.
[94] Nusbaum L, Natour J, Ferraz MB, Goldenberg J. Translation, adaptation and validation of the Roland-Morris Questionnaire-Brazil Roland-Morris. Braz J Med Biol Res 2001;34:203−10.
[95] Osthus H, Cziske R, Jacobi E. Cross-cultural adaptation of a German version of the Oswestry Disability Index and evaluation of its measurement properties. Spine 2006;31:E448−53.

[96] Payares K, Lugo LH, Morales V, Londoño A. Validation in Colombia of the Oswestry disability questionnaire in patients with low back pain. Spine 2011;36:E1730−5.
[97] Pekkanen L, Kautiainen H, Ylinen J, Salo P, Häkkinen A. Reliability and validity study of the Finnish version 2.0 of the Oswestry Disability Index. Spine 2011;36:332−8.
[98] Riecke J, Holzapfel S, Rief W, Lachnit H, Glombiewski JA. Cross-cultural adaption of the German Quebec Back Pain Disability Scale: an exposure-specific measurement for back pain patients. J Pain Res 2016;9:9.
[99] Robinson-Papp J, George MC, Dorfman D, Simpson DM. Barriers to chronic pain measurement: a qualitative study of patient perspectives. Pain Med 2015;16:1256−64.
[100] Rodrigues MF, Michel-Crosato E, Cardoso JR, Traebert J. Psychometric properties and cross-cultural adaptation of the Brazilian Quebec back pain disability scale questionnaire. Spine 2009;34:E459−64.
[101] Sanjaroensuttikul N. The Oswestry low back pain disability questionnaire (version 1.0) Thai version. J Med Assoc Thai 2007;90:1417.
[102] Scharovsky A, Pueyrredón M, Craig D, Rivas ME, Converso G, Pueyrredón JH, et al. Cross-cultural adaptation and validation of the Argentinean version of the Roland-Morris Disability Questionnaire. Spine 2008;33:1391−5.
[103] Suh KT, Kim JI, Lim JM, Goh TS, Lee JS. Validation of the Korean version of the Quebec Back Pain Disability Scale. J spinal Disord Tech 2012;25:447−50.
[104] Tsang RC. Measurement properties of the Hong Kong Chinese version of the Roland-Morris Disability Questionnaire. Hong Kong Physiother J 2004;22:40−9.
[105] Valasek T, Varga PP, Szövérfi Z, Bozsodi A, Klemencsics I, Fekete L, et al. Validation of the Hungarian version of the Roland−Morris Disability Questionnaire. Disabil Rehabil 2015;37(1):86−90.
[106] Valasek T, Varga PP, Szövérfi Z, Kümin M, Fairbank J, Lazary A. Reliability and validity study on the Hungarian versions of the Oswestry Disability Index and the Quebec Back Pain Disability Scale. Eur Spine J 2013;22:1010−8.
[107] van Hooff ML, Spruit M, Fairbank JC, van Limbeek J, Jacobs WC. The Oswestry Disability Index (version 2.1 a): validation of a Dutch language version. Spine 2015;40:E83−90.
[108] Vigatto R, Alexandre NMC, Correa Filho HR. Development of a Brazilian Portuguese version of the oswestry disability index: cross-cultural adaptation, reliability, and validity. Spine 2007;32:481−6.
[109] Vincent JI, MacDermid JC, Grewal R, Sekar VP, Balachandran D. Translation of oswestry disability index into Tamil with cross cultural adaptation and evaluation of reliability and validity (§). Open Orthop J 2014;8:11.
[110] Vogler D, Paillex R, Norberg M, de Goumoens P, Cabri J. Validation transculturelle de l'Oswestry disability index en français Cross-cultural validation of the Oswestry disability index in French. Ann de réadaptation de médecine physique 2008;379−85.
[111] Wei X, Yi H, Wu B, Qi M, Liu X, Chen Z, et al. A valid cross-culturally adapted simplified Chinese version of the Quebec Back Pain Disability Scale. J Clin Epidemiol 2012;65:1321−8.
[112] Wiesinger GF, Nuhr M, Quittan M, Ebenbichler G, Wölfl G, Fialka-Moser V. Cross-cultural adaptation of the Roland-Morris questionnaire for German-speaking patients with low back pain. Spine 1999;24:1099−103.
[113] Yakut E, Düger T, Öksüz Ç, Yörükan S, Üreten K, Turan D, et al. Validation of the Turkish version of the Oswestry Disability Index for patients with low back pain. Spine 2004;29:581−5.
[114] Yi H, Ji X, Wei X, Chen Z, Wang X, Zhu X, et al. Reliability and validity of simplified Chinese version of Roland-Morris questionnaire in evaluating rural and urban patients with low back pain. PLoS one 2012;7(1):e30807.
[115] Yvanes-Thomas M, Calmels P, Béthoux F, Richard A, Nayme P, Payre D, et al. Validity of the French-language version of the Quebec back pain disability scale in low back pain patients in France. Joint Bone Spine 2002;69:397−405.
[116] Algarni A, Ghorbel S, Jones J, Guermazi M. Validation of an Arabic version of the Oswestry index in Saudi Arabia. Ann Phys Rehabil Med 2014;57:653−63.
[117] Bernstein IH, Jaremko ME, Hinkley BS. On the utility of the west haven-yale multidimensional pain inventory. Spine 1995;20:956−63.
[118] Brodke DS, Goz V, Lawrence BD, Spiker WR, Neese A, Hung M. Oswestry Disability Index: a psychometric analysis with 1,610 patients. Spine J 2017;17:321−7.
[119] Brodke DS, Goz V, Voss MW, Lawrence BD, Spiker WR, Man H. PROMIS (R) PF CAT outperforms the ODI and SF-36 physical function domain in spine patients. Spine 2017;42:921−9.
[120] Choi B. Measurement precision for Oswestry Back Pain Disability Questionnaire versus a web-based computer adaptive testing for measuring back pain. J back Musculoskelet Rehabil 2015;28:145−52.
[121] Cook KF, Choi SW, Crane PK, Deyo RA, Johnson KL, Amtmann D. Letting the CAT out of the bag: comparing computer adaptive tests and an eleven-item short form of the Roland-Morris Disability Questionnaire. Spine 2008;33:1378.
[122] Davidson M. Rasch analysis of three versions of the oswestry disability questionnaire. Man Ther 2008;13:222−31.
[123] Davidson M. Rasch analysis of 24-, 18-and 11-item versions of the roland-Morris Disability Questionnaire. Qual Life Res 2009;18:473−81.
[124] Davidson M, Keating JL, Eyres S. A low back-specific version of the SF-36 Physical Functioning scale. Spine 2004;29:586−94.
[125] Exner V, Keel P. Measuring disability of patients with low-back pain−validation of a German version of the Roland & Morris disability questionnaire. Schmerz 2000;14:392−400.
[126] Fisher K, Johnston M. Validation of the Oswestry low back pain disability questionnaire, its sensitivity as a measure of change following treatment and its relationship with other aspects of the chronic pain experience. Physiother Theor Pract 1997;13(1):67−80.
[127] Garratt AM. Rasch analysis of the Roland disability questionnaire. Spine 2003;28:79−84.
[128] Grotle M, Wilkens P, Garratt AM, Scheel I, Storheim K. Which Roland-Morris Disability Questionnaire? Rasch analysis of four different versions tested in a Norwegian population. J Rehabil Med 2013;45:670−7.
[129] Kent P, Grotle M, Dunn KM, Albert HB, Lauridsen HH. Rasch analysis of the 23-item version of the Roland Morris Disability Questionnaire. J Rehabil Med 2015;47:356−64.
[130] Klemenc-Ketiš Z. Disability in patients with chronic non-specific low back pain: validation of the Slovene version of the Oswestry disability index. Slovenian J Public Health 2011;50:87−94.
[131] Küçükdeveci AA, Tennant A, Elhan AH, Niyazoglu H. Validation of the Turkish version of the Roland-Morris Disability Questionnaire for use in low back pain. Spine 2001;26:2738−43.
[132] Lochhead LE, MacMillan PD. Psychometric properties of the Oswestry disability index: Rasch analysis of responses in a work-disabled population. Work 2013;46:67−76.
[133] Magnussen LH, Lygren H, Strand LI, Hagen EM, Breivik K. Reconsidering the Roland-Morris Disability Questionnaire: time for a multidimensional framework? Spine 2015;40:257−63.
[134] Mielenz TJ, Carey TS, Edwards MC. Item response theory analysis of the modified Roland-Morris Disability Questionnaire in a population-based study. Spine 2015;40:E366−71.
[135] Nambi SG. Reliability, validity, sensitivity and specificity of Guajarati version of the Roland-Morris Disability Questionnaire. J Back Musculoskelet Rehabil 2013;26:149−53.
[136] Page SJ, Shawaryn MA, Cernich AN, Linacre JM. Scaling of the revised Oswestry low back pain questionnaire. Arch Phys Med Rehabil 2002;83:1579−84.

[137] Payares K, Lugo LH, Restrepo A. Validation of the Roland Morris Questionnaire in Colombia to evaluate disability in low back pain. Spine 2015;40:1108−14.
[138] Riley JL III, Zawacki TM, Robinson ME, Geisser ME. Empirical test of the factor structure of the West Haven-Yale Multidimensional Pain Inventory. Clin J Pain 1999;15(1):24−30.
[139] Suzukamo Y, Fukuhara S, Kikuchi S, Konno S, Roland M, Iwamoto Y, et al. Validation of the Japanese version of the Roland-Morris disability questionnaire. J Orthop Sci 2003;8:543−8.
[140] Tan G, Jensen MP, Thornby JI, Shanti BF. Validation of the Brief pain inventory for chronic nonmalignant pain. J Pain 2004;5:133−7.
[141] White LJ, Velozo CA. The use of Rasch measurement to improve the Oswestry classification scheme. Arch Phys Med Rehabil 2002; 83:822−31.
[142] Williams RM, Myers AM. Support for a shortened Roland-Morris Disability Questionnaire for patients with acute low back pain. Physiother Can 2001;53:60−6.
[143] Yamato TP, Maher CG, Saragiotto BT, Catley MJ, McAuley JH. The roland−morris disability questionnaire: one or more dimensions? Eur Spine J 2017;26:301−8.
[144] Cook KF, Kallen MA, Amtmann D. Having a fit: impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. Qual Life Res 2009;18: 447−60.
[145] Reise SP. The rediscovery of bifactor measurement models. Multivariate Behav Res 2012;47(5):667−96.
[146] Reise SP, Cook KF, Moore TM. Evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In: Reise SP, Revicki DA, editors. Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment. New York: Routledge; 2014:13−40.
[147] Reise SP, Morizot J, Hays RD. The role of the bifactor model in resolving dimensionality issues in health outcomes measures. Qual Life Res 2007;16:19−31.
[148] De Boeck P, Wilson M. Multidimensional explanatory item response modeling. In: Reise SP, Revicki DA, editors. Handbook of Item Response Theory Modeling. New York: Routledge; 2014: 252−71.
[149] Paap MC, Brouwer D, Glas CA, Monninkhof EM, Forstreuter B, Pieterse ME, et al. The St George's Respiratory Questionnaire revisited: a psychometric evaluation. Qual Life Res 2015;24: 67−79.
[150] van den Berg SM, Paap MC, Derks EM, GROUP Investigators. Using multidimensional modeling to combine self-report symptoms with clinical judgment of schizotypy. Psychiatry Res 2013;206: 75−80.
[151] Fries JF, Cella D, Rose M, Krishnan E, Bruce B. Progress in assessing physical function in arthritis: PROMIS short forms and computerized adaptive testing. J Rheumatol 2009;36:2061−6.
[152] Hays RD, Spritzer KL, Fries JF, Krishnan E. Responsiveness and minimally important difference for the Patient-Reported Outcomes Measurement Information System (PROMIS) 20-item physical functioning short form in a prospective observational study of rheumatoid arthritis. Ann Rheum Dis 2013;74:104−7.
[153] Lee AC, Driban JB, Price LL, Harvey WF, Rodday AM, Wang C. Responsiveness, and minimally important differences for 4 patient-reported outcomes measurement information system short forms: physical function, pain interference, depression, and anxiety in knee osteoarthritis. J Pain 2017;18:1096−110.
[154] Merriwether EN, Rakel BA, Zimmerman MB, Dailey DL, Vance CG, Darghosian L, et al. Reliability and construct validity of the patient-reported outcomes measurement information system (PROMIS) instruments in women with fibromyalgia. Pain Med 2016;18:1485−95.
[155] Wahl E, Gross A, Chernitskiy V, Trupin L, Gensler L, Chaganti K, et al. Validity and responsiveness of a 10-item patient-reported measure of physical function in a rheumatoid arthritis clinic population. Arthritis Care Res 2017;69:338−46.
[156] Chiarotto A, Boers M, Deyo RA, Buchbinder R, Corbin TP, Costa LOP, et al. Core outcome measurement instruments for clinical trials in non-specific low back pain. Pain 2017. https://doi.org/10.1097/j.pain.0000000000001117. [Epub ahead of print].
[157] Magasi S, Ryan G, Revicki D, Lenderking W, Hays RD, Brod M, et al. Content validity of patient-reported outcome measures: perspectives from a PROMIS meeting. Qual Life Res 2012;21:739−46.
[158] Wainer H, Braun HI. Test Validity. Routledge; 2013.